VDJServer: a cloud-based analysis portal and data commons for immune repertoire sequences

Lindsay G. Cowell

Department of Population and Data Sciences

Department of Immunology

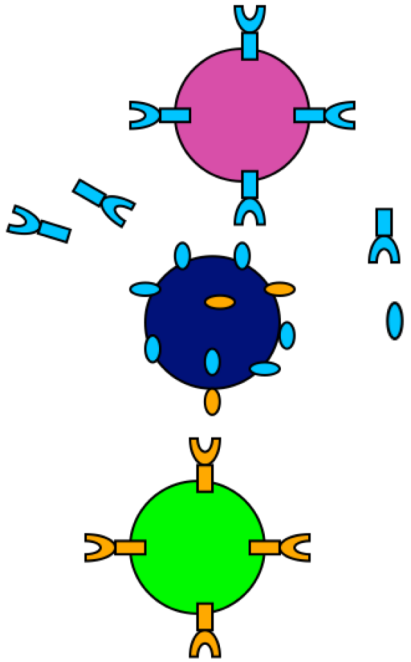
UT Southwestern Medical Center

Lindsay.Cowell@utsouthwestern.edu

Talk Overview

- **Background** – adaptive immune receptor repertoires
- **Data analysis needs** presented by high-throughput sequencing of immune repertoires
- **VDJServer** overview
- Example **clinical research application**

Pathogen and Cancer Detection



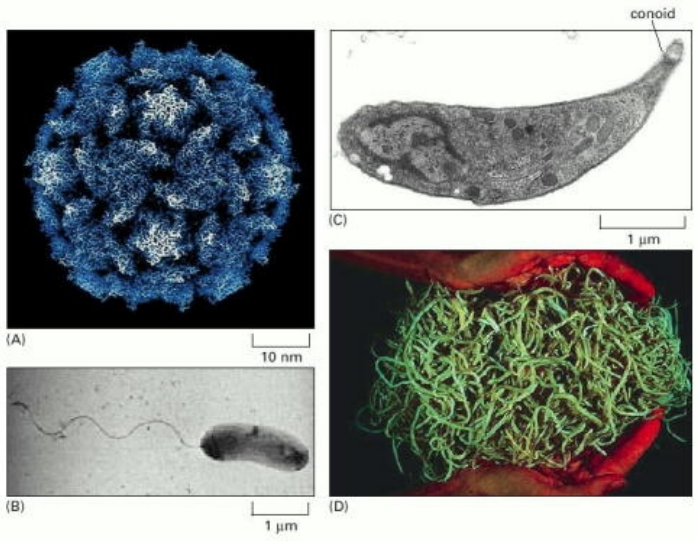
antibody /
antigen receptor

antigen

Lymphocytes express antigen receptors.

Infinite Variety of Antigens

- Highly diverse repertoire of antigens
  - There is a large number of pathogens, and they are phylogenetically diverse.
  - Pathogens and cancers evolve more rapidly than their hosts.
  - They somatically vary their immunogenic molecules.



Molecular Biology of the Cell. 4th edition.
Alberts B, Johnson A, Lewis J, et al.
New York: Garland Science; 2002.

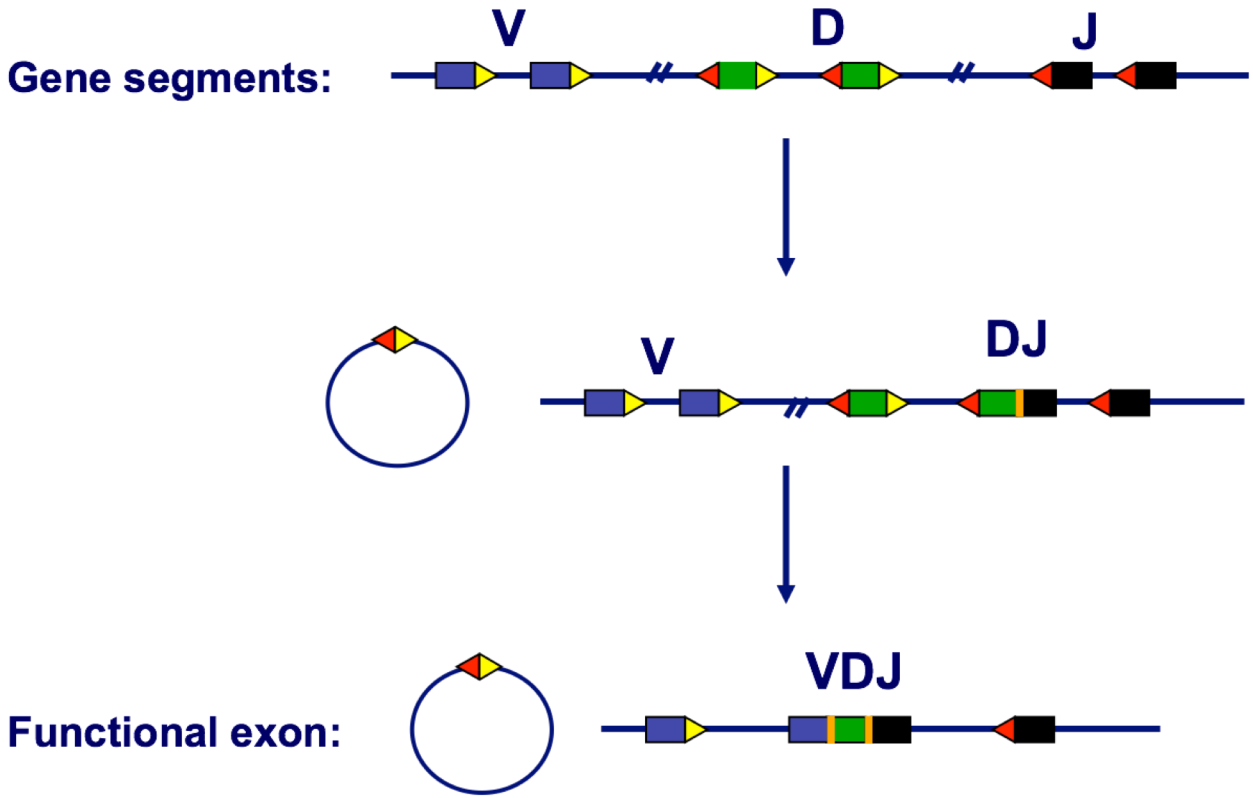
Somatic Generation of Antigen Receptor Genes

- Highly diverse repertoire of antigens
  - There is a large number of pathogens, and they are phylogenetically diverse.
  - Pathogens and cancers evolve more rapidly than their hosts.
  - They somatically vary their immunogenic molecules.

- Highly diverse repertoire of antigen receptors
  - Genes are randomly generated.
  - Genes are subject to subsequent diversification.

Somatic Generation of Genes

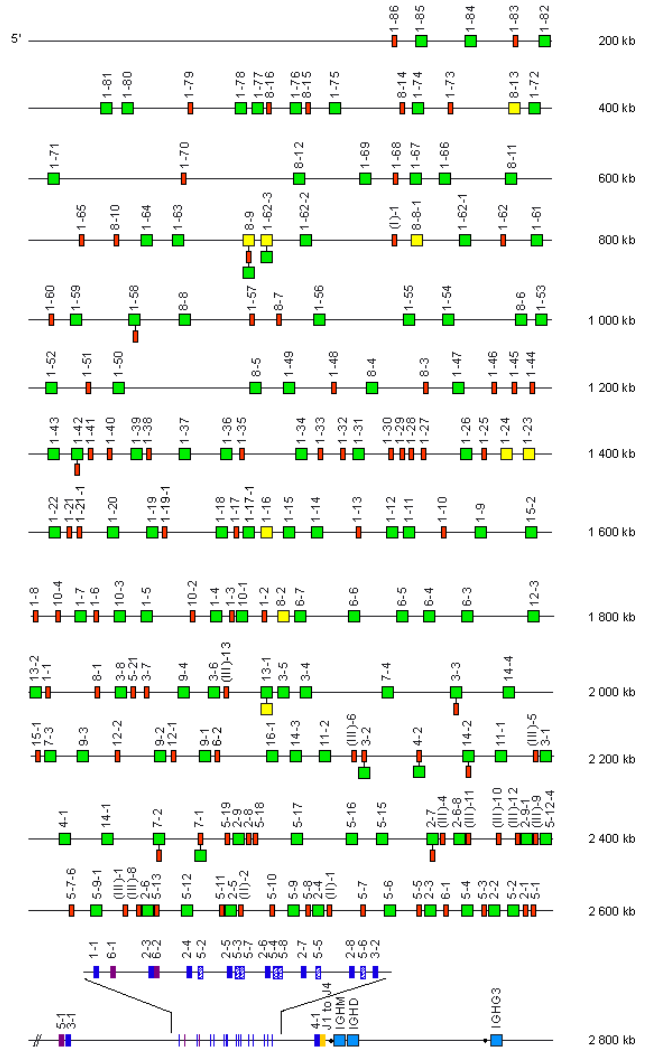
V(D)J Recombination – combinatorial diversity



Somatic Generation of Genes

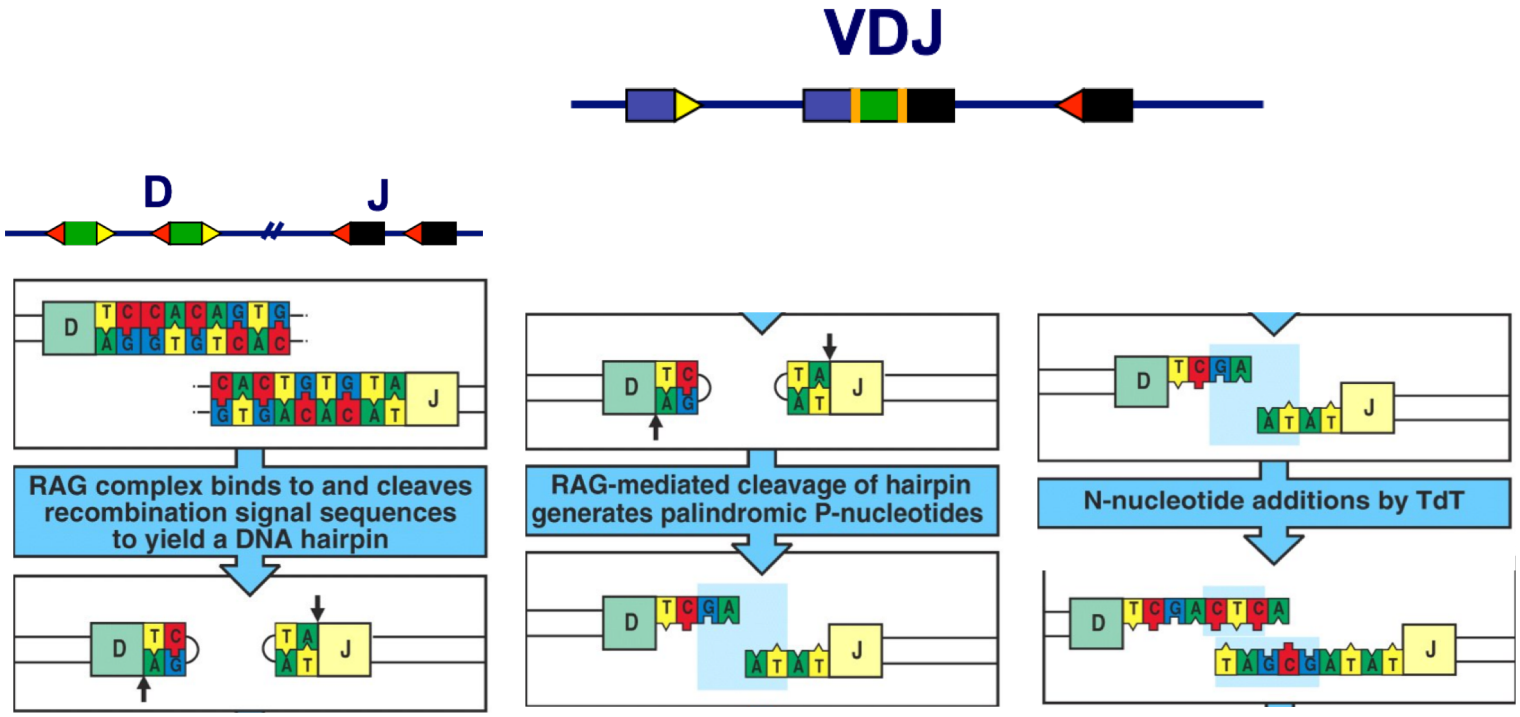
V(D)J Recombination – combinatorial diversity

Mouse IgH locus
3,000 kb
7,676 unique combinations



Somatic Generation of Genes

V(D)J Recombination – junctional diversity



Immunobiology. 6th edition. Janeway, Travers, Walport, Shlomchik

Somatic Generation of Genes

V(D)J Recombination – product

```
input seq.  GTG ACC AGT GCC CAT CCT GAA GAC AGC AGC TTC TAC ATC TGC AGT GCT
       key  VVV VVV VVV VVV VVV VVV VVV VVV VVV VVV VVV VVV VVV VVV VVV VVV
 V 20~1*01  GTG ACC AGT GCC CAT CCT GAA GAC AGC AGC TTC TAC ATC TGC AGT GCT
    D 2~02  ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ..g
  J 2~7*01  ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ...
             81  82  83  84  85  86  87  88  89  90  91  92  93  94  95  96

input seq.  AAA GTA GCG GGA GCT TTC GAC GAG CAG TAC TTC GGG CCG GGC ACC AGG
       key  VVV VDD DDD DDD Dnn nnn nJJ JJJ JJJ JJJ JJJ JJJ JJJ JJJ JJJ JJJ
 V 20~1*01  AgA Ga. ... ... ... ... ... ... ... ... ... ... ... ... ... ...
    D 2~02  ggA cTA GCG GGA Ggg ... ... ... ... ... ... ... ... ... ... ...
  J 2~7*01  ... ... ... ... ..c TcC tAC GAG CAG TAC TTC GGG CCG GGC ACC AGG
             97  98  99 100 101 102 103 104 105 106 107 108 109 110 111 112
```

Bioinformatics. 2006 Feb 15;22(4):438-44. Epub 2005 Dec 15.

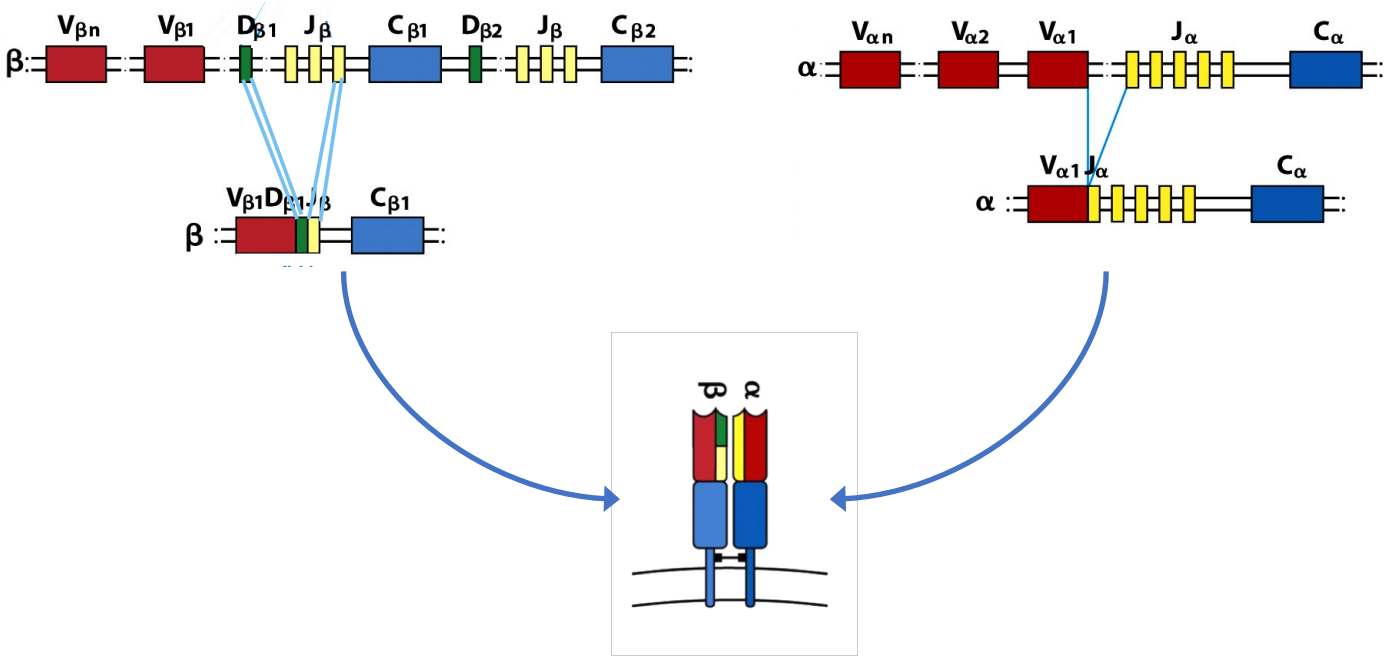
SoDA: implementation of a 3D alignment algorithm for inference of antigen receptor recombinations.

Volpe JM[1], Cowell LG, Kepler TB.

Receptors are Heterodimers

T cell Receptor

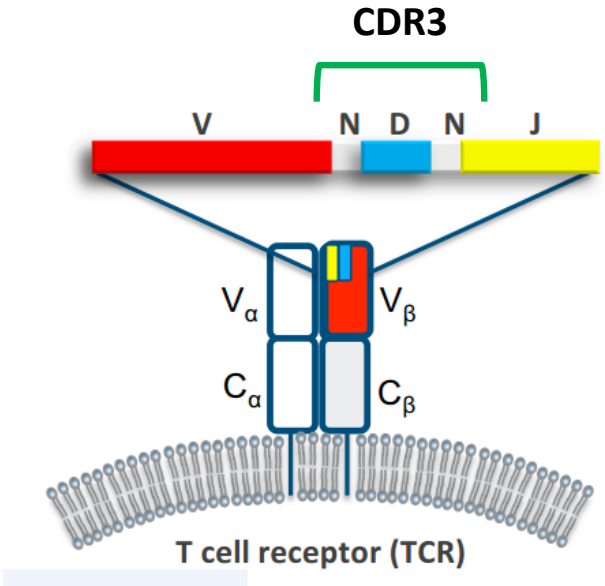
Total Diversity: ~10^{18}



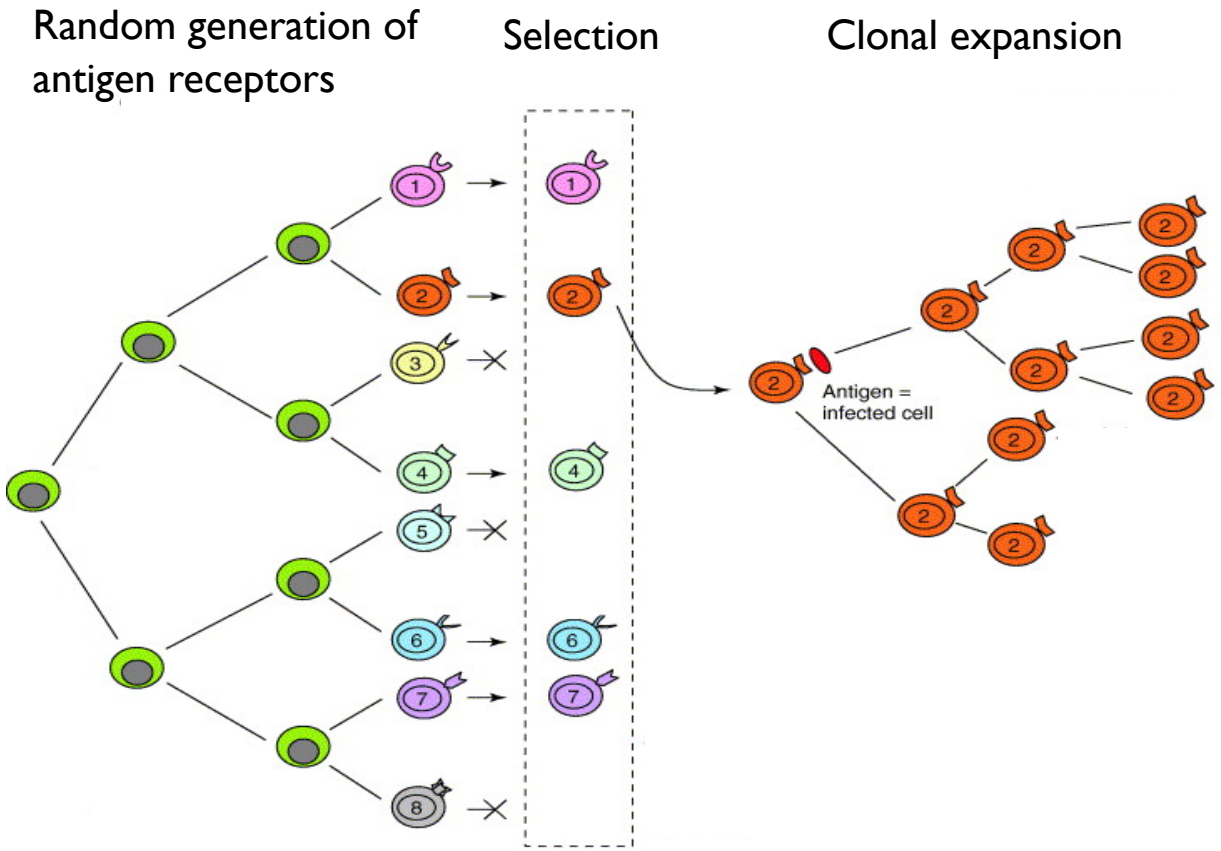
Janeway and Travers remix

Adaptive Immune Receptors

Complementarity Determining Region 3 (CDR3)



Adaptive Immune Receptor Repertoires



Modified from Bergstrom and Antia, 2006. *Trends Ecol Evol*

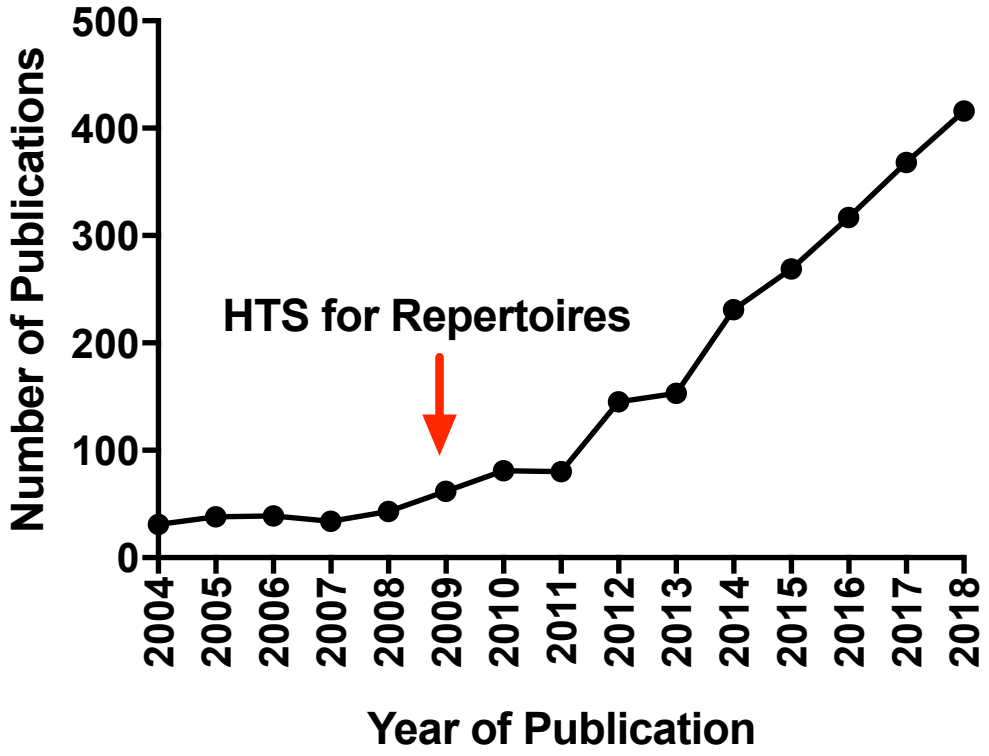
Adaptive Immune Receptor Repertoires

- Adaptive Immune Receptor Repertoire – the full set of functional antigen receptor encoding genes in an individual at a specific time.

- The repertoire shifts in response to antigen exposures.

Repertoire Deep Sequencing

"(repertoire sequencing) AND 2004 [dp]"



Repertoire Profiling

- Deep sequencing of antigen receptor encoding genes
- Enables
  - Enumeration of lymphocyte clones via their CDR3 sequences
  - Quantification of relative clonal abundances
  - Estimation of repertoire diversity
  - Estimation of clonality
  - Tracking of clones over time and between tissues
- Allows inference of population dynamics
  - In migration
  - Clonal expansion

Repertoire Profiling

- Vaccine development
- Identifying the targets of autoimmune responses
- Monitoring patients with autoimmune diseases
- Monitoring patients with immunodeficiencies
- Monitoring patients post transplantation
- Monitoring patients with cancer
- Diagnosing disease?

Informatics Challenges for Repertoire Analysis

- Data management challenges
  - hundreds of millions of reads per run
  - large number of files
  - requires database solutions and automated processing pipelines
- Analysis challenges
  - requires specialized algorithms
  - requires combinations of tools that are not interoperable
  - not deployed for use by experimentalists and clinicians
  - often requires HPC
- Reproducibility challenges
  - provenance,
  - sharing
  - standardized pipelines for reuse

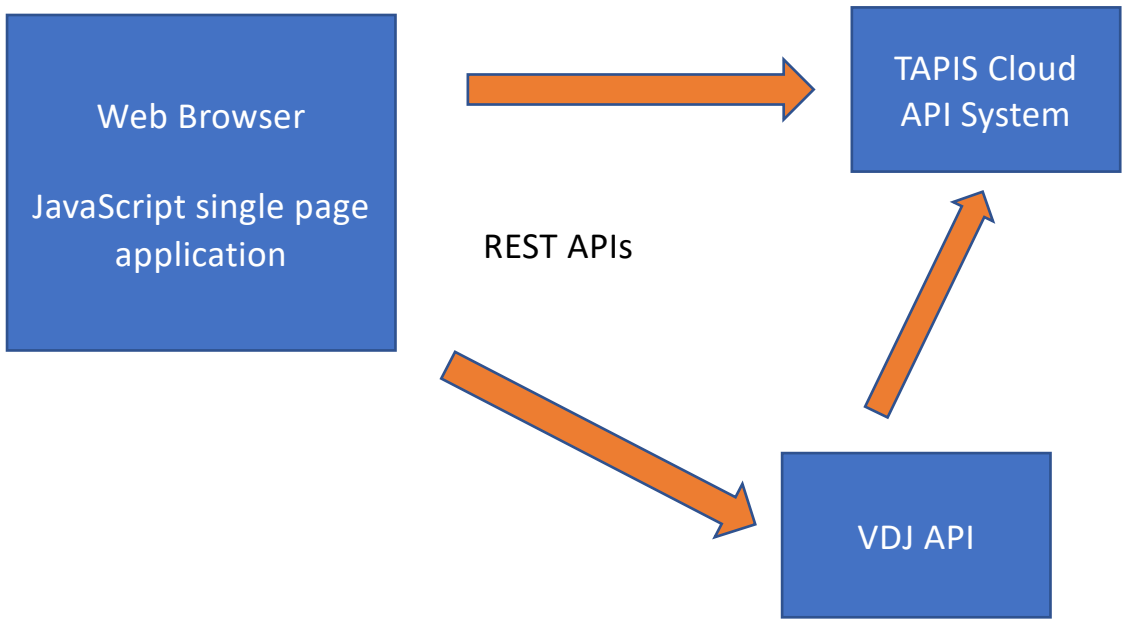
VDJServer – Repertoire Analysis Portal

- data management infrastructure
  - storage with long-term archiving
  - project space that is private, shared, or public
  - study, file, and sample metadata
- sequence processing and analysis portal
  - validated software modules
  - configurable, standardized, reproducible pipelines
  - utilizes the TACC compute resources
  - use metadata to construct comparative analyses
  - automatic collection of analysis provenance

VDJServer – Repertoire Analysis Portal

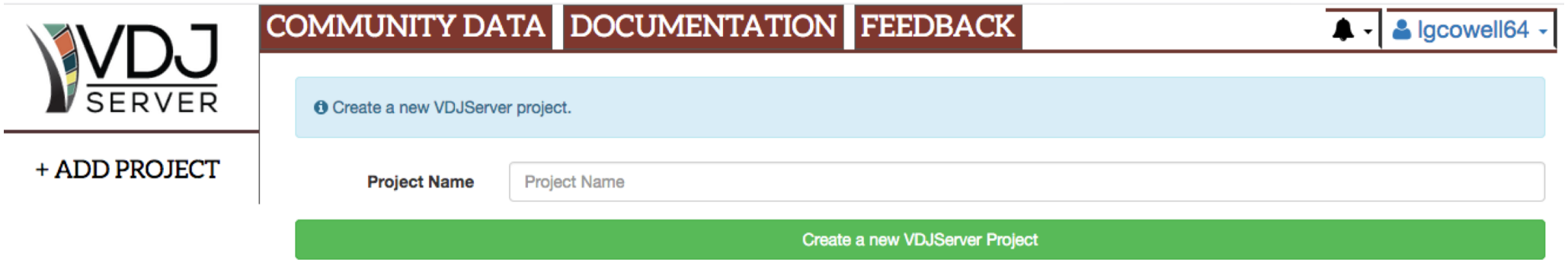
- intuitive user interface and visualization capabilities
  - data management
  - analysis execution
  - interactive visualization to support exploratory analysis
- Community Data Portal
  - public sharing of data
  - public sharing of analyses
  - (will be) query-able

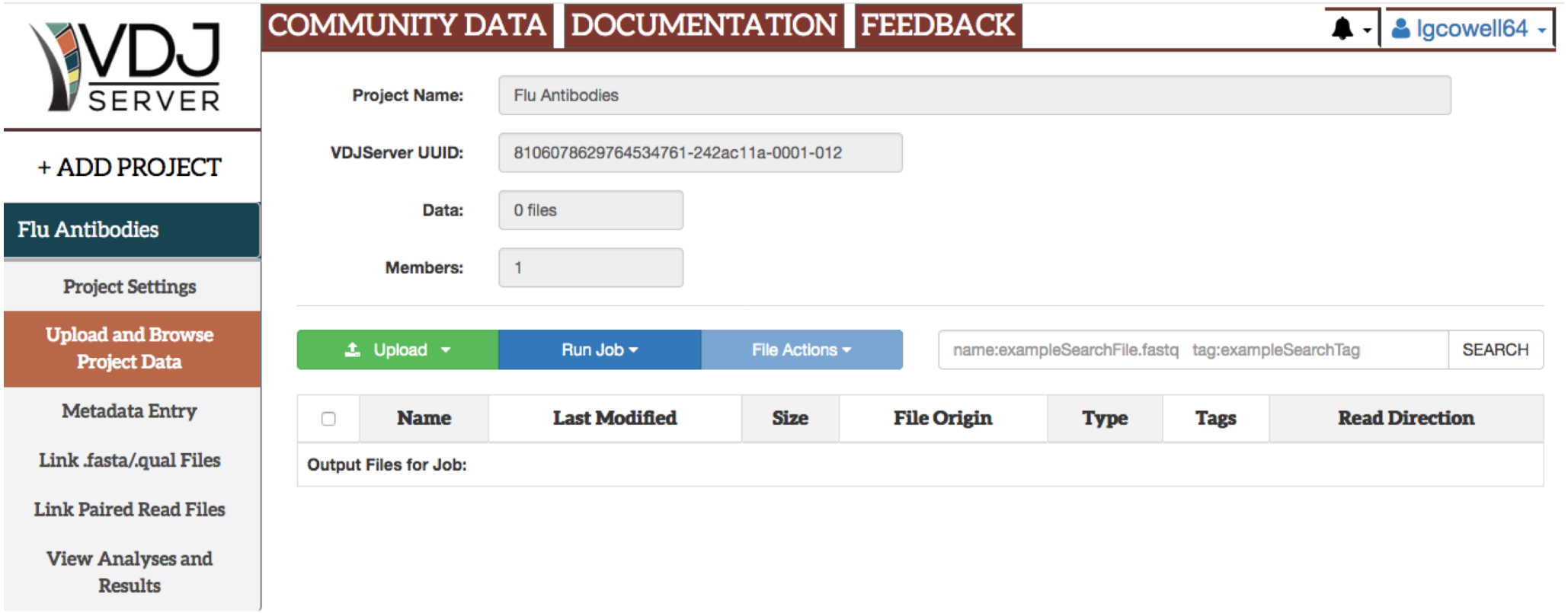
VDJServer Architecture



- Users, files, apps, jobs, notifications, metadata
- Systems:
  - Lonestar5
  - Stampede2

- Pre/Post job tasks
- User account creation, verification and maintenance tasks
- Permissions for project sharing
- Public data access


VDJ SERVER
COMMUNITY DATA
DOCUMENTATION
FEEDBACK
lgcowell64
+ ADD PROJECT
Create a new VDJServer project.
Project Name
Project Name
Create a new VDJServer Project


VDJ SERVER
COMMUNITY DATA
DOCUMENTATION
FEEDBACK
lgcowell64
+ ADD PROJECT
Flu Antibodies
Project Settings
Upload and Browse Project Data
Metadata Entry
Link .fasta/.qual Files
Link Paired Read Files
View Analyses and Results
Project Name:
Flu Antibodies
VDJServer UUID:
8106078629764534761-242ac11a-0001-012
Data:
0 files
Members:
1
Upload
Run Job
File Actions
name:exampleSearchFile.fastq tag:exampleSearchTag
SEARCH
Name
Last Modified
Size
File Origin
Type
Tags
Read Direction
Output Files for Job:

COMMUNITY DATA | DOCUMENTATION | FEEDBACK

scott_ab

+ ADD PROJECT

- Molecular diagnostic test for multiple sclerosis
  - Project Settings
  - Upload and Browse Project Data
  - Metadata Entry
  - Link .fasta/.qual Files
  - Link Paired Read Files
  - View Analyses and Results
- Tissue distribution and clonal diversity of the T and B cell repertoire in type 1 diabetes
- Dynamics of the Cytotoxic T Cell Response to a Model of Acute Viral Infection
- Tumor-infiltrating lymphocytes in colorectal tumors display a diversity of T cell receptor sequences that differ from the T cells in adjacent mucosal tissue
- Antibody repertoire RNA-seq throughout key stages of

Name: Molecular diagnostic test for multiple sclerosis
VDJServer UUID: 30118818971463234 31-242ac11c-0001-012
Data: 12 files
Members: 1

Upload | Run Job | File Actions | name:exampleSearchFile.fastq tag:exampleSearchTag | SEARCH

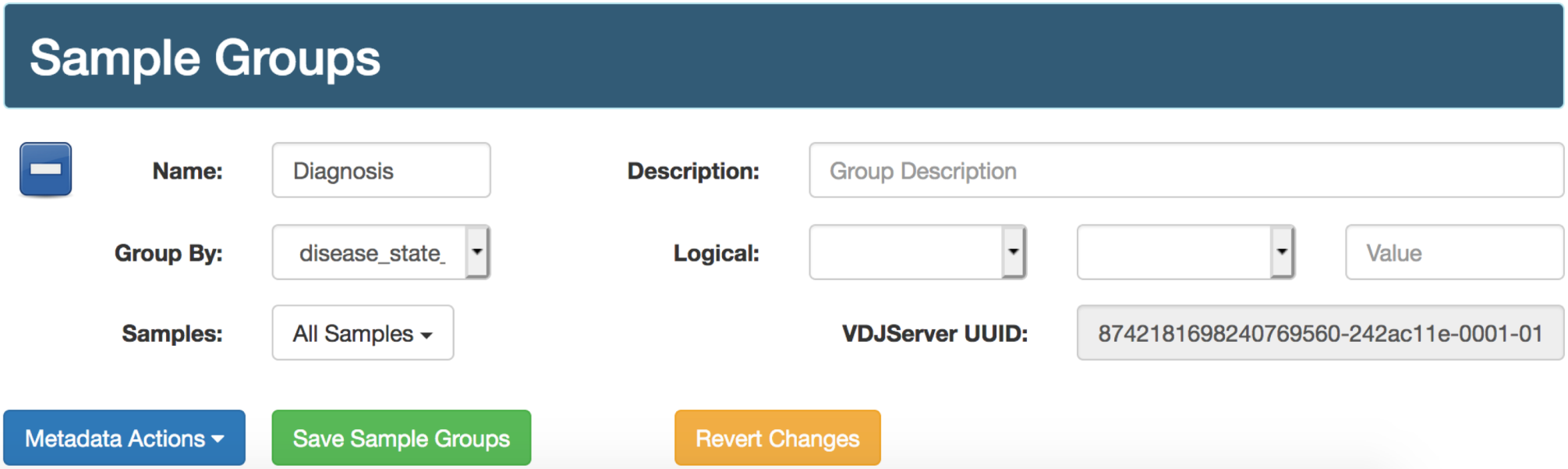
	Name	Last Modified	Size	File Origin	Type	Tags	Read Direction
	Sample00002.fna Sample00002.qual	24-Oct-2017 11:56 am 24-Oct-2017 11:16 am	3.55 MB 9.21 MB	Uploaded File Uploaded File	Single-End Read-Level Data		R
	Sample00003.fna Sample00003.qual	24-Oct-2017 11:56 am 24-Oct-2017 11:16 am	13.74 MB 35.97 MB	Uploaded File Uploaded File	Single-End Read-Level Data		R
	Sample00005.fna Sample00005.qual	24-Oct-2017 11:56 am 24-Oct-2017 11:16 am	8.73 MB 22.72 MB	Uploaded File Uploaded File	Single-End Read-Level Data		R
	Sample00001.fna Sample00001.qual	24-Oct-2017 11:56 am 24-Oct-2017 11:16 am	17.91 MB 46.59 MB	Uploaded File Uploaded File	Single-End Read-Level Data		R

Output Files for Job: VDJPipe pre-processing

	Unique Post-Filter Sequences (Sample00001)	24-Oct-2017 12:02 pm	1.53 MB	Job File	Read-Level Data		
	Unique Post-Filter Sequences (Sample00002)	24-Oct-2017 12:02 pm	857.05 kB	Job File	Read-Level Data		
	Unique Post-Filter Sequences (Sample00005)	24-Oct-2017 12:02 pm	1.52 MB	Job File	Read-Level Data		
	Unique Post-Filter Sequences (Sample00003)	24-Oct-2017 12:01 pm	869.45 kB	Job File	Read-Level Data		

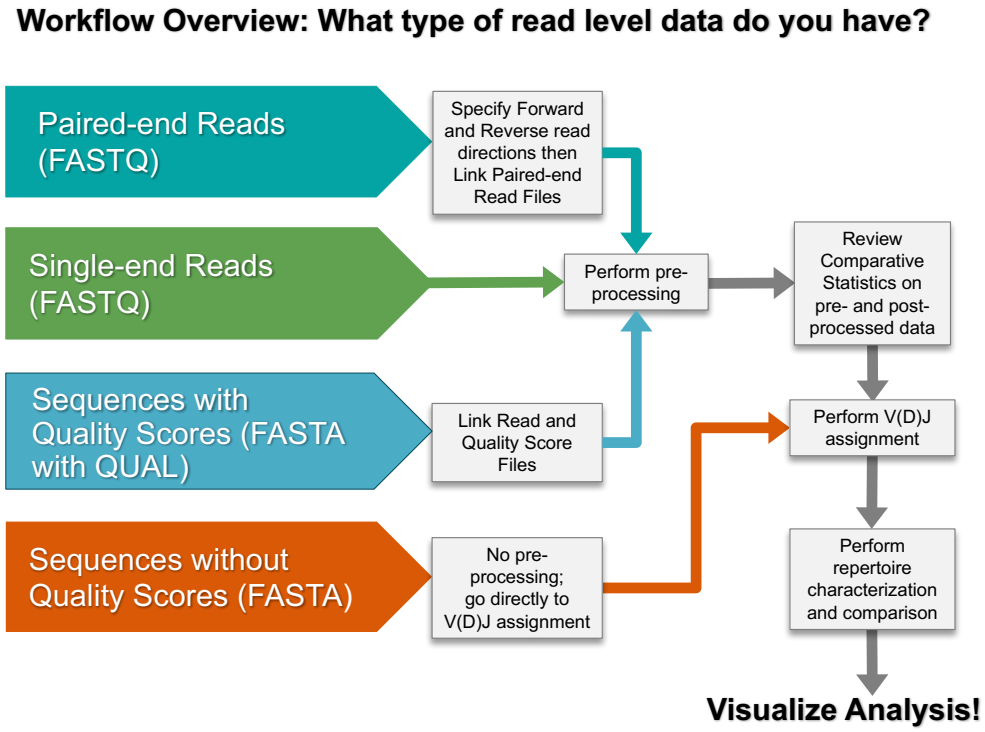
Sample Metadata

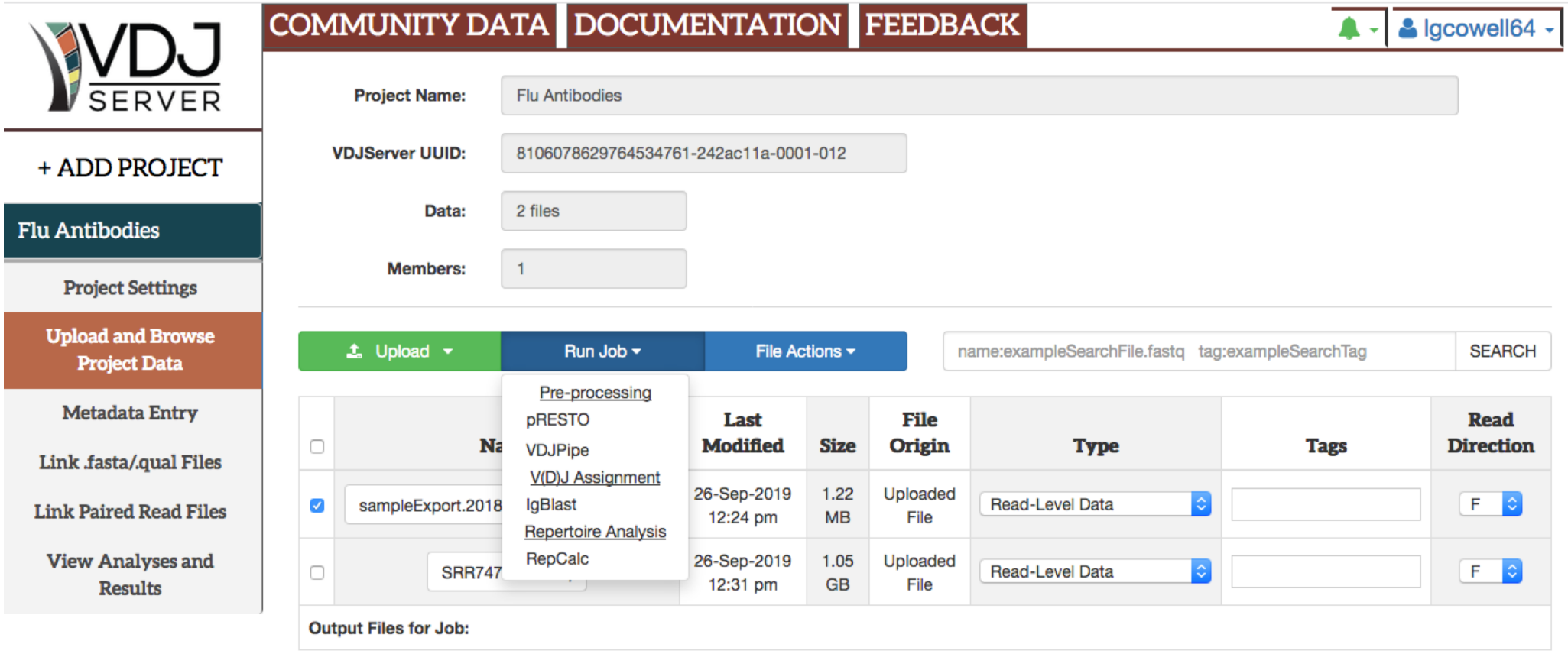
	sample_id	Subject	sample_type	tissue	anatomic_site	disease_state_sample
	O-1B	O-1B	biopsy	ovary	ovary	benign
	O-2B	O-2B	biopsy	ovary	ovary	benign
	O-3B	O-3B	biopsy	ovary	ovary	benign
	O-4B	O-4B	biopsy	ovary	ovary	benign


Sample Groups
Name:
Diagnosis
Description:
Group Description
Group By:
disease_state_
Logical:
Value
Samples:
All Samples
VDJServer UUID:
8742181698240769560-242ac11e-0001-01
Metadata Actions
Save Sample Groups
Revert Changes

VDJServer Basic Analysis Workflow

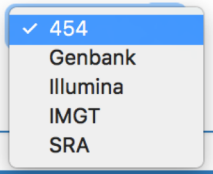
- Upload high-throughput sequencing files
  - ~10GB FASTQ files
- Pre-processing and quality control
- Pre-processing visualizations
- V(D)J Assignment
- Repertoire characterization and comparison
- Analysis visualizations




VDJ SERVER
COMMUNITY DATA
DOCUMENTATION
FEEDBACK
lgcowell64
+ ADD PROJECT
Flu Antibodies
Project Settings
Upload and Browse Project Data
Metadata Entry
Link .fasta/.qual Files
Link Paired Read Files
View Analyses and Results
Project Name:
Flu Antibodies
VDJServer UUID:
8106078629764534761-242ac11a-0001-012
Data:
2 files
Members:
1
Upload
Run Job
File Actions
name:exampleSearchFile.fastq tag:exampleSearchTag
SEARCH
Pre-processing
pRESTO
VDJPipe
V(D)J Assignment
IgBlast
Repertoire Analysis
RepCalc
Last Modified
Size
File Origin
Type
Tags
Read Direction
sampleExport.2018
26-Sep-2019 12:24 pm
1.22 MB
Uploaded File
Read-Level Data
F
SRR747
26-Sep-2019 12:31 pm
1.05 GB
Uploaded File
Read-Level Data
F
Output Files for Job:

Sequence Type

Sequence Type



This module identifies the sequence type used to create input files.

Length/Quality Filter

Minimum Length: 250

Minimum Quality: 20

This module removes reads with a length or contains nucleotides with a quality score below the thresholds provided.

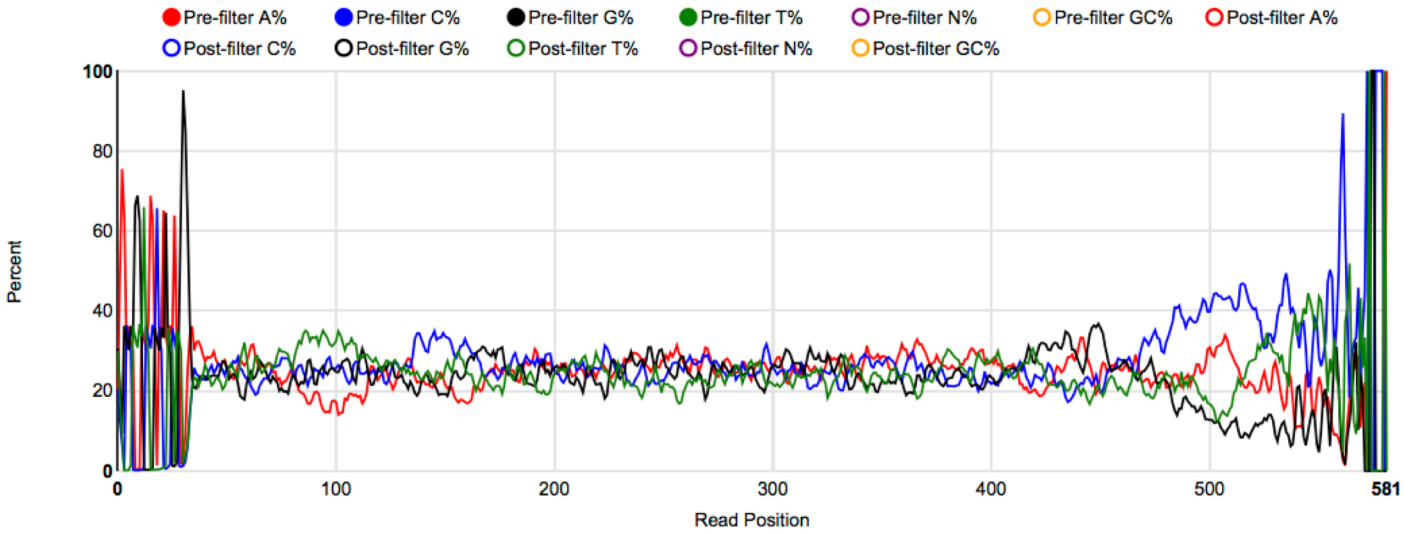
Demultiplex Barcodes

Warning: there is no barcode file available. Please upload barcode file on the project file list and set its type as "Barcode Sequences". Or remove this processing step from the workflow if it is not needed.

Pre-processing Visualizations

A. Nucleotide distribution
B. GC% content
C. Sequence length
D. Average sequence quality
E. Detailed sequence quality at each nucleotide position

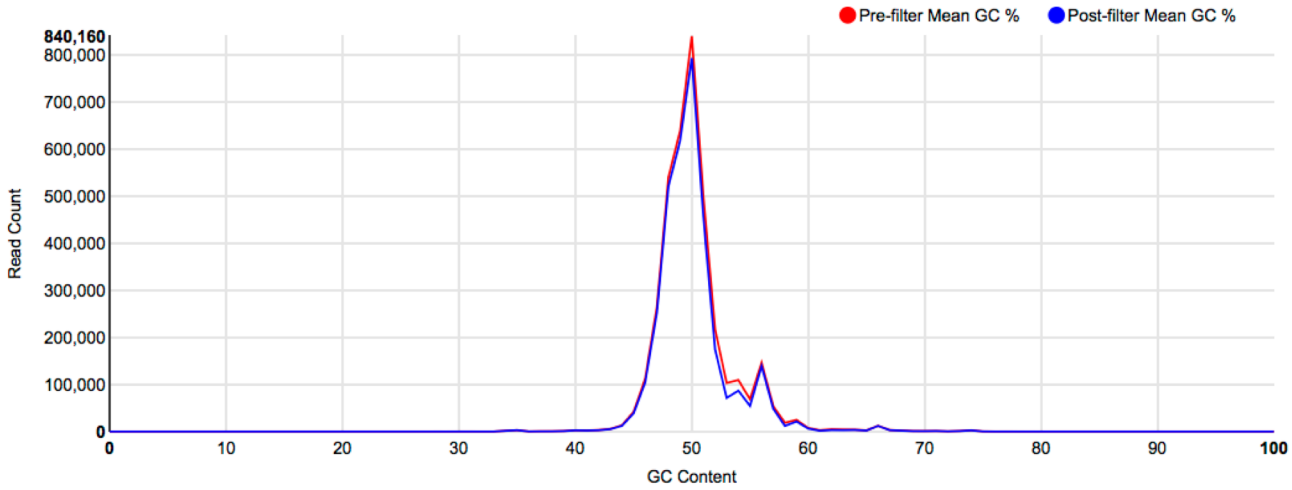
Graphs can show both pre- and post-processing statistics for comparison.



Pre-processing Visualizations

A. Nucleotide distribution
B. GC% content
C. Sequence length
D. Average sequence quality
E. Detailed sequence quality at each nucleotide position

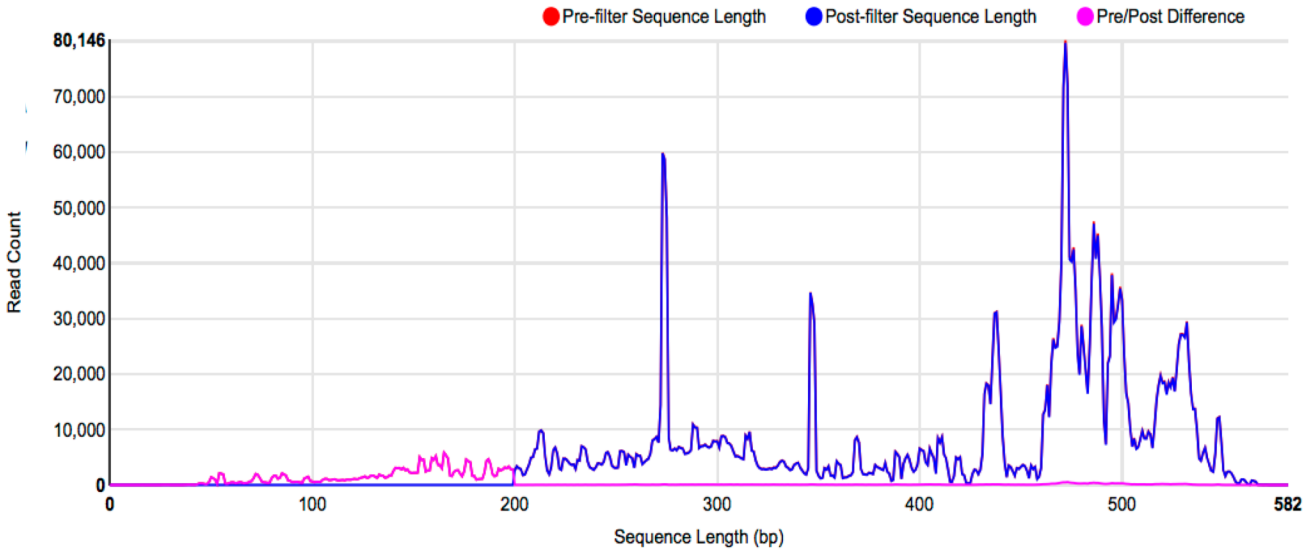
Graphs can show both pre- and post-processing statistics for comparison.



Pre-processing Visualizations

A. Nucleotide distribution
B. GC% content
C. Sequence length
D. Average sequence quality
E. Detailed sequence quality at each nucleotide position

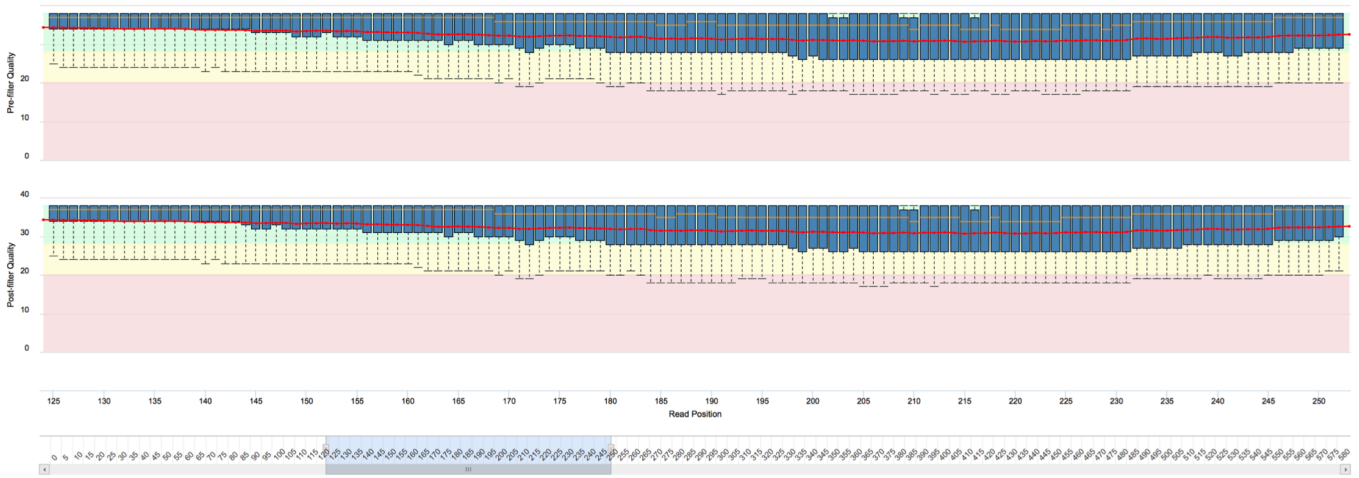
Graphs can show both pre- and post-processing statistics for comparison.



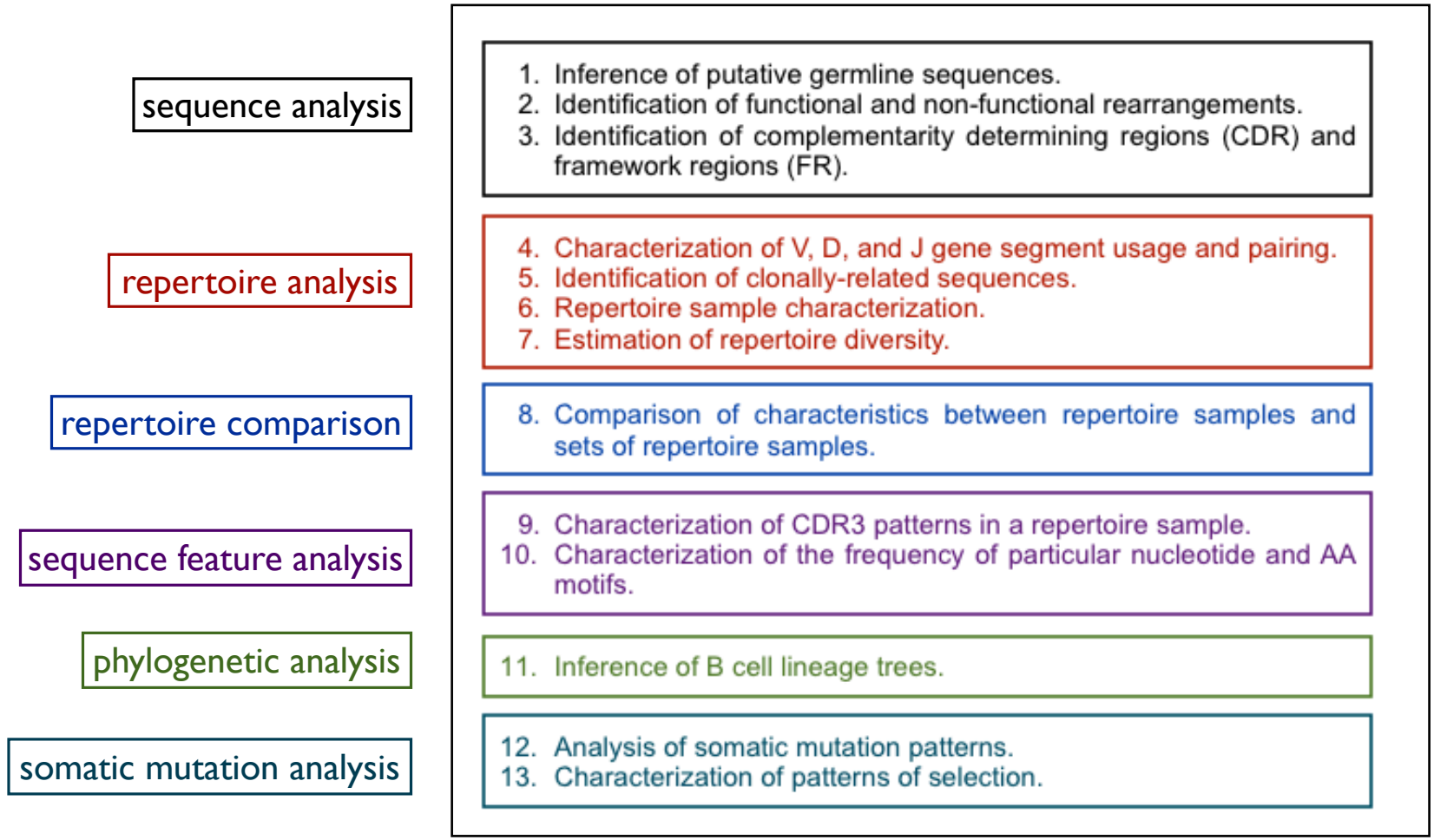
Pre-processing Visualizations

A. Nucleotide distribution
B. GC% content
C. Sequence length
D. Average sequence quality
E. Detailed sequence quality at each nucleotide position

Graphs can show both pre- and post-processing statistics for comparison.



Repertoire analysis modules



Analysis Charts

Chart: Relative Gene Segment Usage [Show Chart] [Download Data]

Files: None selected **Samples:** O-13M **Sample Groups:** None selected

- Diagnosis (sample.disease_state_sample = malignant)
- Diagnosis (sample.disease_state_sample = benign)
- Diagnosis (sample.disease_state_sample = normal)

Chart: Gene Segment Usage [Show Chart] [Download Data]

Files: None selected **Samples:** O-13M **Sample Groups:**

Chart: CDR3 (AA) Length Histogram [Show Chart] [Download Data]

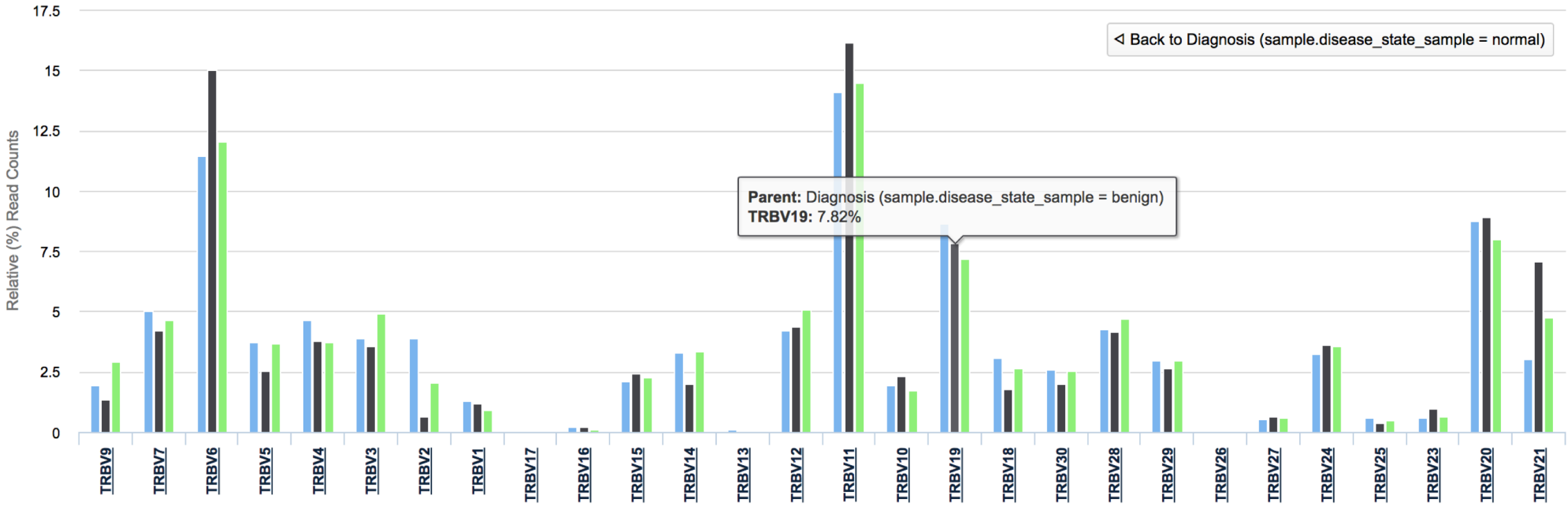
Files: None selected **Samples:** O-13M **Sample Groups:** None selected

Chart: CDR3 (NA) Length Histogram [Show Chart] [Download Data]

Files: None selected **Samples:** O-13M **Sample Groups:** None selected

Chart: Relative Gene Segment Usage | Hide Chart | Download Chart | Download Data

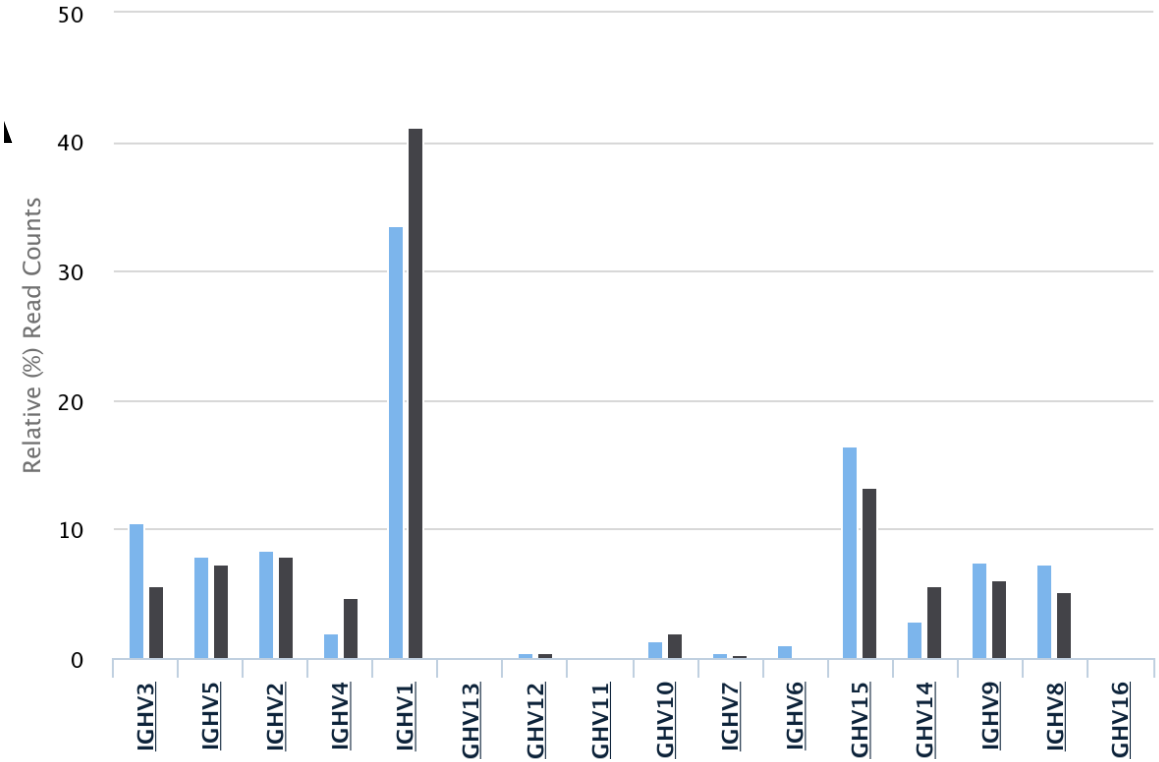
Files: None selected ▾ **Samples:** None selected ▾ **Sample Groups:** All selected (3) ▾



Analysis Visualizations

A. Gene segment usage
B. Length histograms
C. Clonal abundance
D. Cumulative clonality
E. Diversity profiles
F. Selection quantification

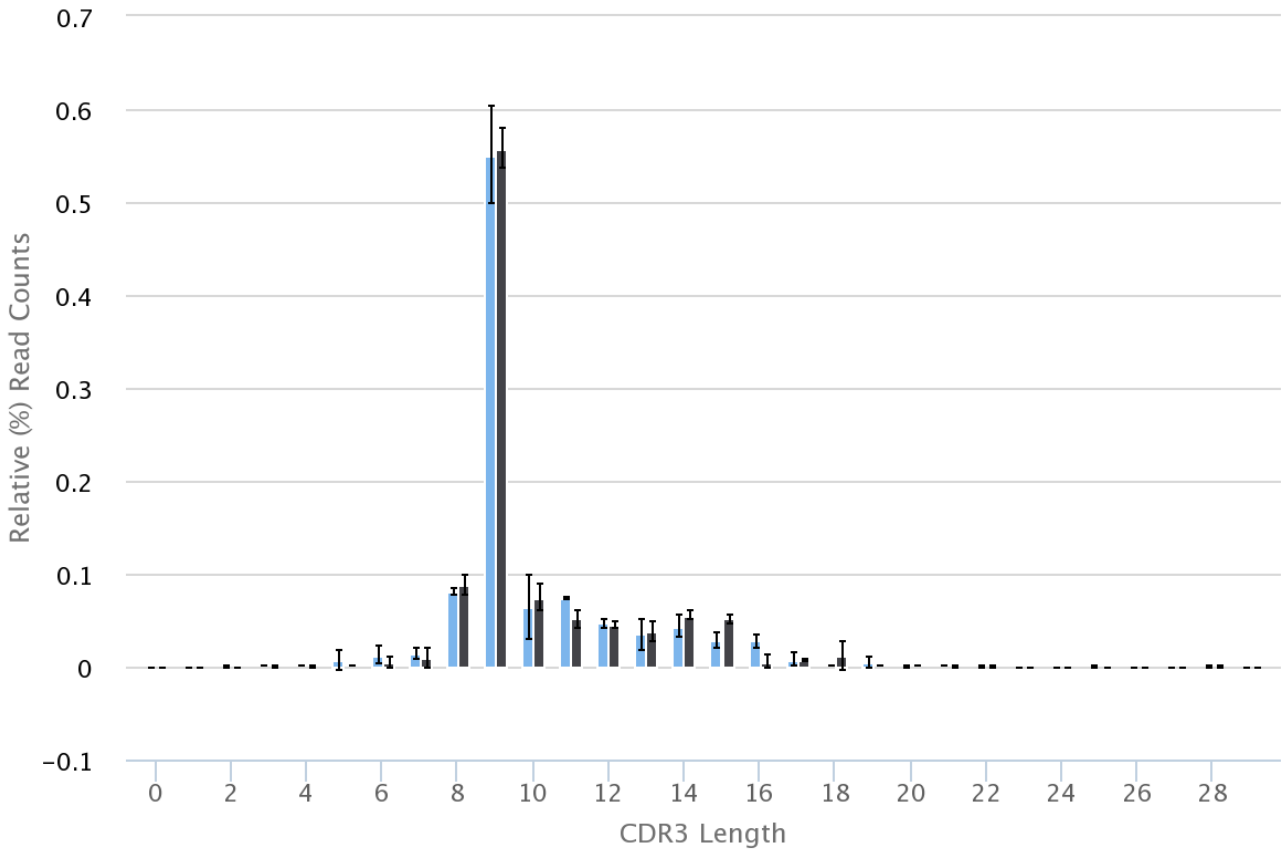
Graphs are interactive allowing repertoires and groups to be dynamically added/removed for comparison.



Analysis Visualizations

A. Gene segment usage
B. Length histograms
C. Clonal abundance
D. Cumulative clonality
E. Diversity profiles
F. Selection quantification

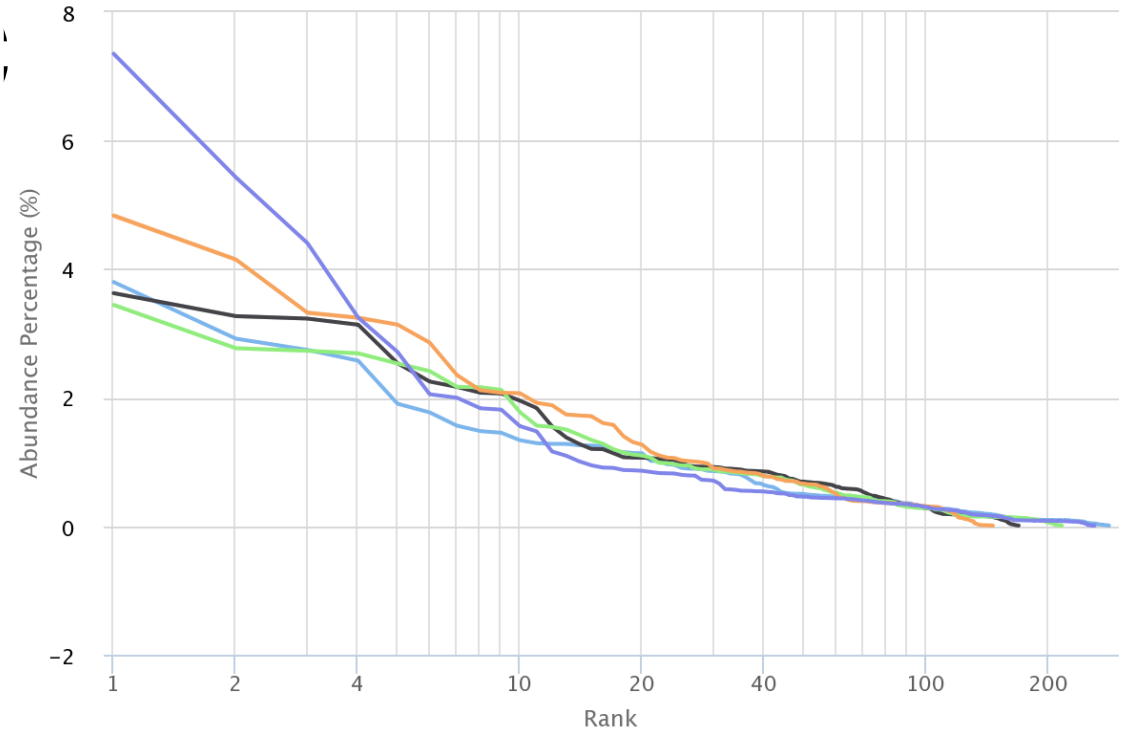
Graphs are interactive allowing repertoires and groups to be dynamically added/removed for comparison.



Analysis Visualizations

A. Gene segment usage
B. Length histograms
C. Clonal abundance
D. Cumulative clonality
E. Diversity profiles
F. Selection quantification

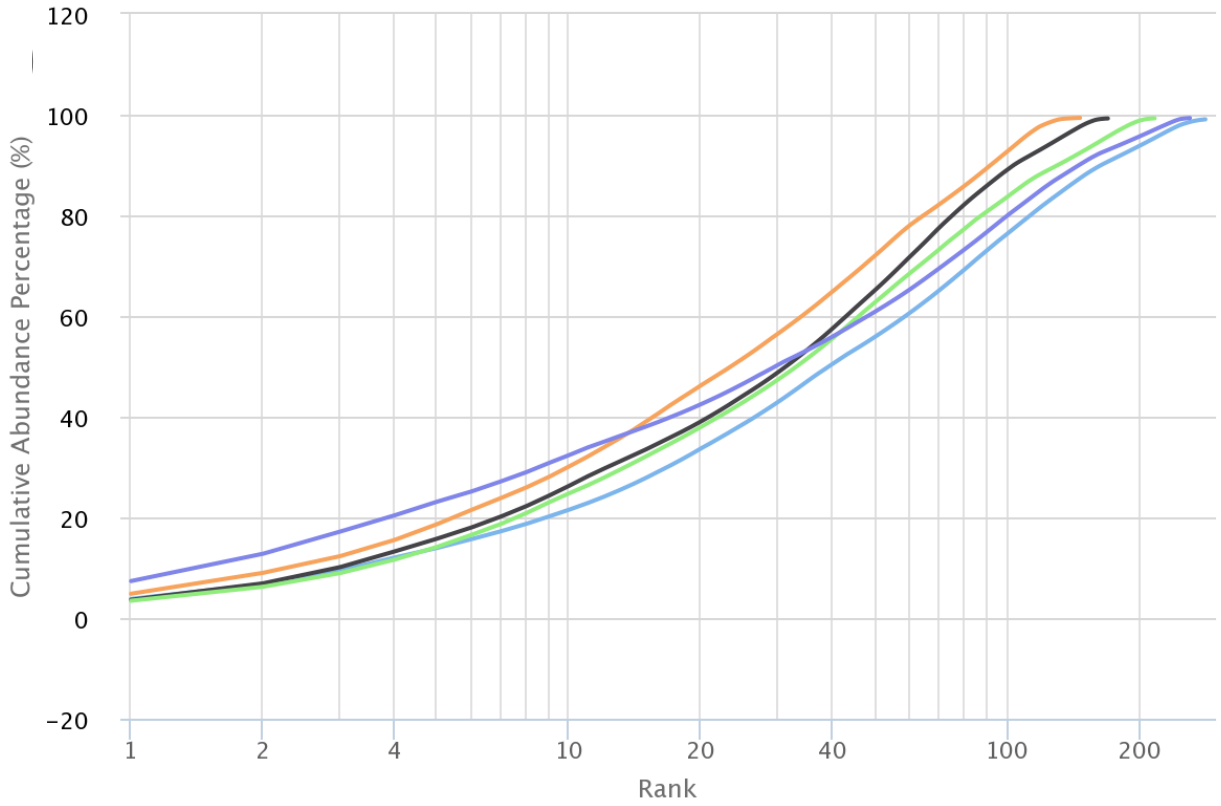
Graphs are interactive allowing repertoires and groups to be dynamically added/removed for comparison.



Analysis Visualizations

A. Gene segment usage
B. Length histograms
C. Clonal abundance
D. Cumulative clonality
E. Diversity profiles
F. Selection quantification

Graphs are interactive allowing repertoires and groups to be dynamically added/removed for comparison.



VDJServer Community Data Portal

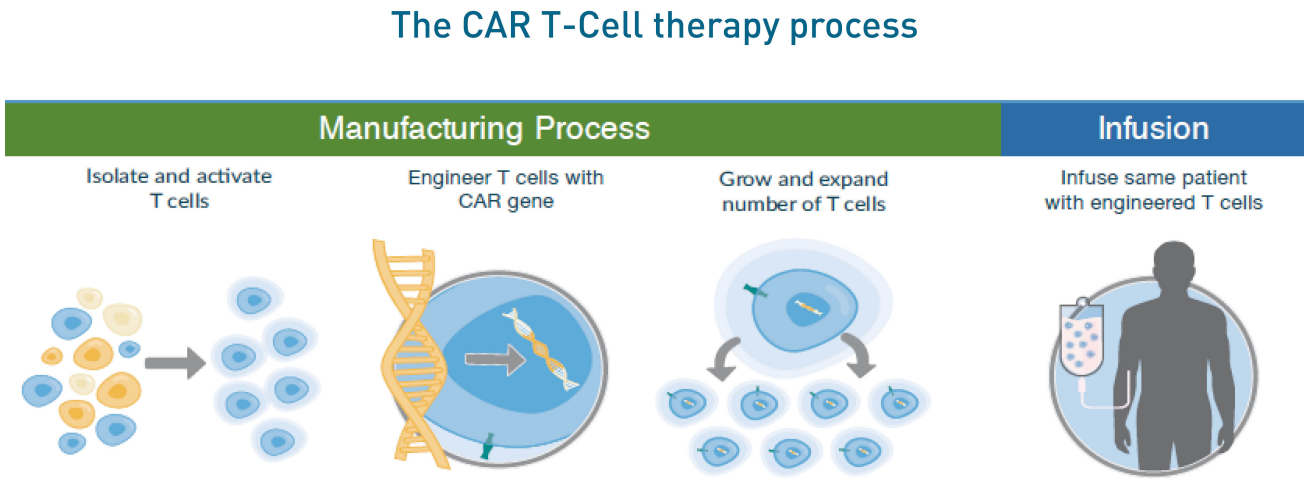
- Users can publish their projects making the data, metadata, analyses, and visualizations available to the public.

  - Planned: hundreds of studies, billions of rearrangement records, TB scale
  - Database managed by the TACC database group
  - Query through the Tapis v3 metadata API

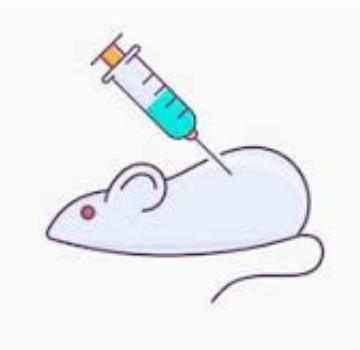
  - Users can perform comparative analysis between their private data and data queried from the AIRR Data Commons

Example Clinical Application – CAR T cells

- Marco L. Davila, MD., PhD., Moffitt Cancer Center



Example Clinical Application – CAR T cells



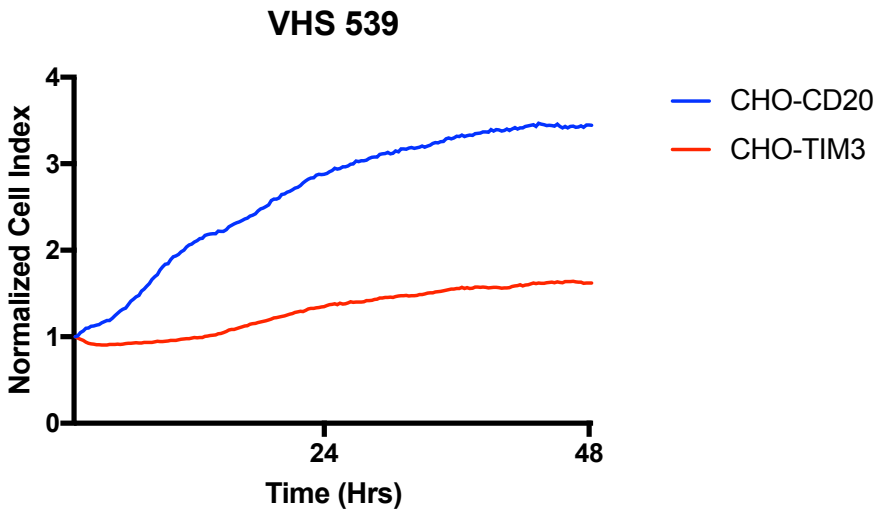
Chinese Hamster Ovary Cells

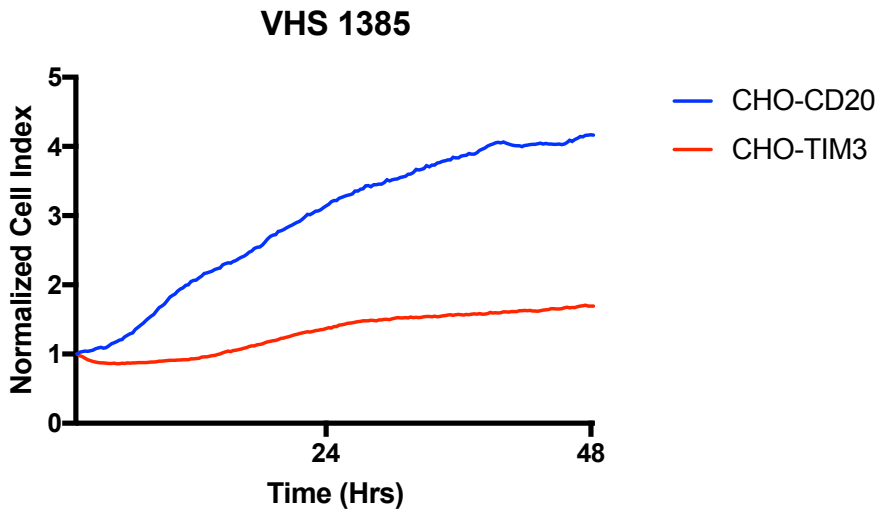
Chinese Hamster Ovary Cells expressing TIM3

Example Clinical Application – CAR T cells

- Deep sequencing of B cell receptor repertoire from both groups
- Use VDJServer to:
  - Identify highly clonally expanded B cell receptors unique to the TIM3 group
- Engineer CAR T cells and assess in an in vitro killing assay

Example Clinical Application – CAR T cells





Acknowledgements

- **TACC**
  - John Fonner
  - Walter Scarborough
  - Steve Mock
  - Matt Stelmazek
  - Chris Jordan
  - Joe Stubbs
- Richard Scheuermann, JCVI
- Nancy Monson, UTSW
- **Cowell Research Group**
  - Scott Christley
  - Ini Toby
  - William Rounds
  - Min Kim
  - Eddie Salinas
  - Mikhail Levin

Funders
- NIAID
- EU Commission
- UT Southwestern